Université Moulay Ismaïl

École Nationale Supérieure d'Arts et Métiers - Meknès

PROJET

CHAID

Chi-squared Automatic Interaction Detection

Méthode d'arbre de décision pour l'analyse de données catégorielles

---

Réalisé par : Marouane Majidi
Walid El Majdi
Chadia Naim

Encadré par : M. Mohamed Hosni

Année Universitaire : 2025/2026

Résumé

Ce rapport présente une étude approfondie de la méthode CHAID (CHi-square Automatic Interaction Detector), une technique d'arbre de décision développée par G.V. Kass en 1980 pour l'analyse de données catégorielles. Nous retraçons d'abord l'évolution historique des méthodes d'arbres, depuis AID jusqu'à CHAID, en explicitant les limitations qui ont motivé chaque développement. Nous exposons ensuite en détail les fondements statistiques de CHAID, notamment le test du χ^2, puis nous décrivons l'algorithme complet avec ses variantes (avec et sans correction de Bonferroni).

Table des matières

1 Historique : de AID à THAID puis CHAID

1.1 Contexte et motivation

Les méthodes d'arbres de décision trouvent leur origine dans l'analyse de données d'enquêtes sociales des années 1950-1960. À cette époque, avec la multiplication des enquêtes empiriques en sciences sociales, les chercheurs se sont heurtés aux limites des modèles linéaires classiques.

Limitations du modèle linéaire. Dans le modèle linéaire, les effets des variables explicatives sont essentiellement additifs : l'effet d'une variable est supposé indépendant des valeurs prises par les autres variables. Or, comme le soulignent Morgan et Sonquist (1963) :

> *« Il existe deux raisons puissantes de croire qu'il est erroné de supposer que les diverses influences sont additives. Premièrement, de nombreux effets d'interaction puissants sont déjà connus — l'éducation supérieure aide davantage un homme qu'une femme en matière de revenus [...] Deuxièmement, les classifications mesurées ne sont que des proxies pour plus d'un construit [...] Nous pouvons avoir des effets d'interaction non pas parce que le monde est plein d'interactions, mais parce que nos variables doivent interagir pour produire les construits théoriques qui importent réellement. »*

Cette nécessité de détecter et modéliser les interactions complexes a motivé le développement des méthodes d'arbres.

1.2 AID (Automatic Interaction Detection)

1.2.1 Principe général

La méthode AID, proposée par Morgan et Sonquist (1963) et popularisée par le programme informatique développé à Ann Arbor (Sonquist, Baker et Morgan, 1971), est conçue pour les **variables dépendantes continues**.

Mécanisme de partitionnement. AID procède par divisions binaires successives :

i. À chaque nœud, on considère toutes les variables explicatives disponibles

ii. Pour chaque variable, on cherche la meilleure dichotomie possible

iii. On sélectionne la variable et la dichotomie qui maximisent la réduction de la somme des carrés des résidus

Critère statistique. Le critère utilisé dans AID est la maximisation de la réduction de la somme des carrés intra-groupes (WSS — Within Sum of Squares) :

$$\text{WSS} = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2$$

où $\bar{y}_j$ est la moyenne des observations dans le nœud j.

Cela équivaut à maximiser le coefficient $\eta^2 = \text{BSS/TSS}$, où :

$$\text{TSS} = \sum_{j=1}^{g} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2 \quad \text{(somme totale des carrés)}$$

$$\text{BSS} = \text{TSS} - \text{WSS} \quad \text{(somme inter-groupes)}$$

1.2.2 Types de variables explicatives

AID distingue deux types de prédicteurs :

— **Prédicteurs monotoniques** : variables ordinales dont les catégories sont ordonnées. Seules les catégories contiguës peuvent être groupées ensemble.

— **Prédicteurs libres (free)** : variables nominales pures, sans ordre naturel. Toutes les combinaisons de catégories sont permises.

1.2.3 Limites de AID

Plusieurs critiques majeures ont été formulées contre AID :

1. **Absence de prise en compte de la variabilité d'échantillonnage.** Comme le note Bishop, Fienberg et Holland (1975, p.360) : *« AID a de sérieuses limitations car il ne prend jamais vraiment en compte la variabilité d'échantillonnage inhérente aux données. »*
2. **Biais en faveur des prédicteurs avec beaucoup de catégories.** La procédure de sélection favorise les variables ayant plus de modalités, puisque le critère de maximisation s'étend sur plus de possibilités.
3. **Distribution du point de scission.** Hawkins (1975) a montré que sous l'hypothèse nulle d'un groupe homogène, la position du point de scission suit une distribution en forme de U, conduisant à des divisions déséquilibrées sans justification statistique.
4. **Absence de tests de significativité.** Aucun cadre statistique formel ne permet de juger si une division est significative ou due au hasard.

1.3 THAID (THeta AID)

1.3.1 Extension aux variables catégorielles

Face à la nécessité d'analyser des **variables dépendantes catégorielles**, Messenger et Mandell (1972) puis Morgan et Messenger (1973) ont proposé THAID.

Critère thêta. Le critère thêta consiste à maximiser la somme du nombre d'observations dans chaque catégorie modale :

$$\theta = \sum_{j=1}^{g} n_j^{\text{modal}}$$

où n_j^{modal} est le nombre d'observations dans la catégorie la plus fréquente du groupe j.

Remarque 1. *Le critère thêta correspond au taux d'erreur de classification : on prédit pour chaque groupe la catégorie modale, et on minimise le nombre d'erreurs.*

1.3.2 Limitations de THAID

1. **Manque de fondement théorique.** Contrairement au critère AID (qui repose sur l'analyse de variance), le comportement théorique du critère thêta reste mal compris.
2. **Conservation des limites de AID.** THAID hérite des problèmes de AID concernant les tests de significativité et le biais selon le nombre de catégories.
3. **Divisions binaires uniquement.** Comme AID, THAID impose des divisions binaires, ce qui peut être inefficace quand plusieurs groupes naturels existent.

1.4 Motivations pour CHAID

Face à ces limitations, Kass (1980) propose CHAID avec les innovations suivantes :

1. **Cadre statistique formel** : utilisation systématique de tests de significativité (test du χ^2)
2. **Divisions multi-branches** : au lieu d'imposer des dichotomies, CHAID permet de créer autant de groupes que nécessaire
3. **Correction du biais** : l'utilisation de p-values avec correction de Bonferroni neutralise le biais en faveur des variables avec beaucoup de modalités
4. **Nouveau type de prédicteur** : introduction des prédicteurs « flottants » pour gérer les valeurs manquantes
5. **Heuristique de fusion** : algorithme efficace pour trouver des partitions optimales sans énumération exhaustive

2 CHAID : définition générale et intuition

2.1 Définition formelle

Définition 1 (CHAID)**.** *CHAID (CHi-square Automatic Interaction Detector) est une technique d'arbre de décision qui partitionne récursivement un ensemble de données en sous-groupes mutuellement exclusifs et exhaustifs, de manière à maximiser la dépendance statistique entre les variables explicatives et une variable dépendante catégorielle. Le critère de partition repose sur le test d'indépendance du χ^2, et la méthode permet des divisions multi-branches.*

2.2 Caractéristiques principales

Variables considérées.

— **Variable dépendante** : catégorielle avec $d \geq 2$ modalités

— **Variables explicatives** : catégorielles (nominales, ordinales ou continues discrétisées)

Remarque 2. *Bien que Kass (1980) se concentre sur les variables dépendantes catégorielles, CHAID est souvent implémenté avec une option pour les variables continues, utilisant alors le test F au lieu du χ^2.*

Objectif statistique. À chaque nœud, CHAID cherche à :

1. Pour chaque prédicteur, déterminer le **meilleur regroupement optimal** de ses catégories
2. Sélectionner le prédicteur dont le regroupement optimal est **le plus significativement associé** à la variable dépendante
3. Partitionner les données selon ce regroupement optimal

2.3 Intuition : pourquoi le test du χ^2 ?

Le test du χ^2 mesure l'écart entre les distributions observées et les distributions attendues sous l'hypothèse d'indépendance.

Lien avec la dépendance statistique.

— Si deux variables sont **indépendantes**, la distribution de Y est la même dans tous les groupes définis par X

— Si elles sont **dépendantes**, les distributions de Y varient selon les groupes

— Le χ^2 quantifie cette variation : plus il est élevé, plus la dépendance est forte

Avantage sur les mesures de pureté. Contrairement à CART ou ID3 qui utilisent des mesures de pureté (Gini, entropie) visant à créer des nœuds homogènes, CHAID se concentre sur la **force de l'association**. Cela correspond mieux à l'objectif exploratoire original : détecter les interactions significatives.

2.4 Pourquoi des divisions multi-branches ?

Efficacité descriptive. Lorsqu'une variable a naturellement plusieurs groupes distincts, forcer des divisions binaires :

1. Complexifie inutilement l'arbre
2. Rend l'interprétation moins directe
3. Peut masquer la structure réelle des données

3 Le test du χ^2 dans CHAID

3.1 Principe du test du χ^2

3.1.1 Hypothèses statistiques

Le test du χ^2 d'indépendance teste les hypothèses suivantes :

H_0 : La variable explicative X et la variable dépendante Y sont indépendantes
H_1 : X et Y sont dépendantes (associées)

Remarque 3. *Rejeter H_0 signifie que la distribution de Y varie significativement selon les modalités de X, ce qui justifie d'utiliser X pour partitionner les données.*

3.1.2 Table de contingence

Considérons une variable explicative X avec c catégories et une variable dépendante Y avec d catégories. La table de contingence est une matrice $c \times d$ où :

$$n_{ij} = \text{nombre d'observations dans la catégorie } i \text{ de } X \text{ et la catégorie } j \text{ de } Y$$

Les totaux marginaux sont :

$$n_{i\cdot} = \sum_{j=1}^{d} n_{ij} \quad \text{(total de la ligne } i\text{)}$$

$$n_{\cdot j} = \sum_{i=1}^{c} n_{ij} \quad \text{(total de la colonne } j\text{)}$$

$$n = \sum_{i=1}^{c} \sum_{j=1}^{d} n_{ij} \quad \text{(total général)}$$

3.2 Calcul mathématique du χ^2

3.2.1 Fréquences attendues sous H_0

Sous l'hypothèse d'indépendance, la fréquence attendue dans la cellule (i, j) est :

$$e_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

Interprétation. Si X et Y sont indépendantes, la proportion d'observations de la catégorie j de Y devrait être la même dans tous les groupes définis par X. Cette proportion globale est $n_{\cdot j}/n$, et appliquée au groupe i (de taille $n_{i\cdot}$), elle donne e_{ij}.

3.2.2 Statistique du χ^2

La statistique de test est :

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{d} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

Cette formule mesure l'écart quadratique pondéré entre observations et attentes.

3.2.3 Degrés de liberté

Les degrés de liberté sont :

$$\text{ddl} = (c - 1)(d - 1)$$

Justification. Une fois fixés les totaux marginaux (qui résument l'information sous H_0), seules $(c - 1)(d - 1)$ cellules sont libres de varier ; les autres sont déterminées par contrainte.

3.2.4 Calcul de la p-value

Sous H_0, la statistique χ^2 suit asymptotiquement une loi du χ^2 à $(c-1)(d-1)$ degrés de liberté. La p-value est :

$$p = P(\chi^2_{\text{ddl}} \geq \chi^2_{\text{obs}})$$

Remarque 4. *Une p-value faible (typiquement < 0.05) conduit à rejeter H_0 : on conclut à une dépendance significative entre X et Y.*

3.3 Interprétation statistique dans CHAID

3.3.1 Lien entre χ^2, p-value et dépendance

1. χ^2 **élevé** $\Rightarrow$ écarts importants entre observations et attentes $\Rightarrow$ forte dépendance
2. **p-value faible** $\Rightarrow$ χ^2 élevé est peu probable sous H_0 $\Rightarrow$ rejet de l'indépendance
3. **Dans CHAID** : on choisit le prédicteur avec la p-value la plus faible (dépendance la plus significative)

3.3.2 Seuils de décision

CHAID utilise des seuils α (typiquement 0.01, 0.05 ou 0.10) pour :

— **Fusion** : fusionner deux catégories si $p > \alpha_{\text{fusion}}$

— **Division** : séparer un groupe fusionné si $p < \alpha_{\text{division}}$

— **Sélection** : choisir une variable pour partitionner si $p < \alpha_{\text{split}}$

Remarque 5. *Il est essentiel que $\alpha_{division} < \alpha_{fusion}$ pour assurer la convergence de l'algorithme et éviter les boucles infinies.*

3.4 Exemple numérique détaillé

Considérons un exemple simple tiré de l'analyse de Ritschard (2013) sur les étudiants. Nous examinons la relation entre le **type de diplôme secondaire** (variable X) et la **situation après la première année** (variable Y).

3.4.1 Table de contingence initiale (extraite)

Simplifions à 2 catégories pour X :

— Catégorie 1 : diplôme technique (9 étudiants)

— Catégorie 2 : sans diplôme (12 étudiants)

Et 3 catégories pour Y : éliminé, redoublant, réussi.

	Éliminé	**Redoublant**	**Réussi**	**Total**
Technique (1)	5	0	4	9
Sans diplôme (2)	6	0	6	12
Total	11	0	10	21

3.4.2 Calcul des fréquences attendues

Pour la cellule (1, éliminé) :

$$e_{1,\text{élim}} = \frac{9 \times 11}{21} = \frac{99}{21} = 4.714$$

Pour la cellule (1, réussi) :

$$e_{1,\text{réus}} = \frac{9 \times 10}{21} = \frac{90}{21} = 4.286$$

Pour la cellule (2, éliminé) :

$$e_{2,\text{élim}} = \frac{12 \times 11}{21} = \frac{132}{21} = 6.286$$

Pour la cellule (2, réussi) :

$$e_{2,\text{réus}} = \frac{12 \times 10}{21} = \frac{120}{21} = 5.714$$

Les cellules « redoublant » ont $e_{ij} = 0$ puisque le total de colonne est 0.

3.4.3 Calcul de la statistique χ^2

$$\chi^2 = \frac{(5 - 4.714)^2}{4.714} + \frac{(4 - 4.286)^2}{4.286} + \frac{(6 - 6.286)^2}{6.286} + \frac{(6 - 5.714)^2}{5.714}$$

$$= \frac{0.082}{4.714} + \frac{0.082}{4.286} + \frac{0.082}{6.286} + \frac{0.082}{5.714}$$

$$\approx 0.017 + 0.019 + 0.013 + 0.014 = 0.063$$

3.4.4 Degrés de liberté et p-value

Degrés de liberté : $(c - 1)(d - 1) = (2 - 1)(3 - 1) = 2$
La p-value associée à $\chi^2 = 0.063$ avec 2 ddl est très élevée (environ 0.969 ou 96.9%).

3.4.5 Interprétation dans CHAID

Cette p-value très élevée signifie que les distributions de la situation académique sont très similaires entre les étudiants avec diplôme technique et ceux sans diplôme. Par conséquent, CHAID fusionnerait ces deux catégories, considérant qu'elles ne méritent pas d'être distinguées pour prédire la variable dépendante.

4 L'algorithme CHAID

4.1 Types de variables explicatives

CHAID distingue trois types de prédicteurs, chacun imposant des contraintes différentes sur les regroupements possibles.

4.1.1 Variables nominales (free predictors)

Définition 2 (Prédicteur nominal)**.** *Un prédicteur nominal est une variable catégorielle dont les modalités n'ont pas d'ordre naturel (par exemple : couleur, nationalité, type de diplôme).*

Contrainte de regroupement. Pour un prédicteur nominal, **toute combinaison** de catégories peut être fusionnée. Si le prédicteur a c catégories, il existe un très grand nombre de partitions possibles.

Nombre de partitions. Le nombre de façons de partitionner c catégories en g groupes est donné par le nombre de Stirling de seconde espèce :

$$S(c, g) = \frac{1}{g!} \sum_{i=0}^{g-1} (-1)^i \binom{g}{i} (g - i)^c$$

Le nombre total de partitions (tous nombres de groupes confondus) est le nombre de Bell :

$$B(c) = \sum_{g=1}^{c} S(c, g)$$

Exemple 1. *Pour $c = 4$ catégories $\{a, b, c, d\}$, il y a $B(4) = 15$ partitions possibles, dont $S(4, 2) = 7$ partitions en 2 groupes :*

$$\{a\}\{bcd\},\ \{b\}\{acd\},\ \{c\}\{abd\},\ \{d\}\{abc\},$$
$$\{ab\}\{cd\},\ \{ac\}\{bd\},\ \{ad\}\{bc\}$$

4.1.2 Variables ordinales (monotonic predictors)

Définition 3 (Prédicteur ordinal)**.** *Un prédicteur ordinal est une variable catégorielle dont les modalités possèdent un ordre naturel (par exemple : niveau d'éducation, tranches d'âge, année).*

Contrainte de regroupement. Pour un prédicteur ordinal, seules les **catégories contiguës** peuvent être fusionnées. Cela préserve l'ordre sous-jacent.

Nombre de partitions. Le nombre de façons de partitionner c catégories ordonnées en g groupes contigus est :

$$\binom{c - 1}{g - 1}$$

Exemple 2. *Pour $c = 4$ catégories ordonnées $a < b < c < d$, il y a 3 partitions en 2 groupes :*

$$\{a\}\{bcd\},\ \{ab\}\{cd\},\ \{abc\}\{d\}$$

4.1.3 Variables flottantes (floating predictors)

Définition 4 (Prédicteur flottant)**.** *Un prédicteur flottant est une variable ordinale avec une modalité spéciale (typiquement « valeur manquante » ou « non applicable ») qui ne peut être positionnée naturellement sur l'échelle ordinale.*

Motivation. Les valeurs manquantes sont fréquentes dans les données d'enquête. Plutôt que d'imputer ou de supprimer ces observations, CHAID crée une catégorie flottante qui peut se combiner librement avec n'importe quel autre groupe.

Contrainte de regroupement.

- Les catégories ordinales suivent les règles des prédicteurs ordinaux
- La catégorie flottante peut :
  - Rester seule
  - Se fusionner avec n'importe quelle catégorie ou groupe de catégories

Nombre de partitions. Pour $c-1$ catégories ordinales plus 1 catégorie flottante, le nombre de partitions en g groupes est :

$$\binom{c-2}{g-2} + g\binom{c-2}{g-1} = \frac{g-1+g(c-g)}{c-1}\binom{c-1}{g-1}$$

Exemple 3. *Pour 3 catégories ordonnées $a < b < c$ et une catégorie flottante f, il y a 5 partitions en 2 groupes :*

$$\{af\}\{bc\},\ \{a\}\{bcf\},\ \{abf\}\{c\},\ \{ab\}\{cf\},\ \{abc\}\{f\}$$

4.2 CHAID sans ajustement de Bonferroni

Nous présentons d'abord l'algorithme de base, sans correction pour tests multiples. Cette version illustre la logique fondamentale de CHAID.

4.2.1 Vue d'ensemble de l'algorithme

CHAID procède en 5 étapes principales :

1. **Fusion (Merging)** : Pour chaque prédicteur, regrouper itérativement les catégories similaires (Minimisation du χ^2)
2. **Division (Splitting)** : Vérifier si certains regroupements peuvent être défusionnés (Maximisation du χ^2)
3. **Évaluation** : Calculer la significativité de chaque prédicteur optimal
4. **Sélection** : Choisir le prédicteur le plus significatif et partitionner les données (Maximisation du χ^2)
5. **Récursion** : Répéter le processus sur chaque nœud fils

4.2.2 Étape 1 : Fusion des catégories (Merging)

Objectif. Réduire le nombre de catégories d'un prédicteur en fusionnant celles qui ont des distributions similaires de la variable dépendante.

Procédure détaillée. **Étape 1.1.** Pour un prédicteur avec c catégories, construire la table de contingence $c \times d$ avec la variable dépendante.

Étape 1.2. Calculer le χ^2 pour chaque paire de catégories autorisée (selon le type de prédicteur), en considérant la sous-table $2 \times d$ correspondante.

Étape 1.3. Identifier la paire avec le χ^2 le plus faible.

— **Logique :** On cherche ici la similarité maximale entre catégories. En sélectionnant la différence la moins significative (le χ^2 minimal), on identifie les groupes qui sont suffisamment proches pour être fusionnés.

Étape 1.4. Si cette p-value dépasse le seuil α_{fusion} (typiquement 0.05), fusionner les deux catégories et traiter le résultat comme une catégorie composée unique.

Étape 1.5. Répéter les étapes 1.2 à 1.4 jusqu'à ce que toutes les paires restantes aient un χ^2 significatif (p-value $\leq \alpha_{\text{fusion}}$).

Justification statistique. Fusionner deux catégories non significativement différentes :

— Réduit la complexité sans perte d'information substantielle

— Améliore la stabilité statistique en augmentant les effectifs par groupe

— Évite le sur-ajustement dû à des distinctions non significatives

Partition asymptotique du χ^2. Kass s'appuie sur un résultat de Kendall et Stuart (1961) : la statistique χ^2 totale d'une table $c \times d$ peut être décomposée de manière asymptotiquement valide en :

— Une composante χ^2_{paire} pour la sous-table $2 \times d$ considérée

— Une composante résiduelle pour la table réduite après fusion

Cette propriété justifie de traiter χ^2_{paire} comme une statistique du χ^2 à $d-1$ degrés de liberté.

4.2.3 Étape 2 : Vérification des divisions (Splitting)

Objectif. S'assurer qu'aucun groupe composé de 3 catégories ou plus ne cache une division significative.

Procédure. Pour chaque catégorie composée formée de $k \geq 3$ catégories originales :

1. Examiner toutes les dichotomies possibles (respectant les contraintes du type de prédicteur)
2. Calculer le χ^2 de la division la plus significative.
   — **Logique :** Contrairement à l'étape de fusion, on cherche ici la divergence maximale. En identifiant la division avec le χ^2 le plus élevé, on détermine s'il existe une raison statistique forte de rompre le regroupement établi.
3. Si la p-value est inférieure à α_{division} (typiquement $0.049 < 0.05$), implémenter la division
4. Retourner à l'étape de fusion

Condition de convergence. Pour que l'algorithme converge, il faut : $\alpha_{\text{division}} < \alpha_{\text{fusion}}$ Sinon, on pourrait entrer dans un cycle infini fusion-division.

Remarque 6. *En pratique, les divisions sont rares. L'algorithme de fusion construit généralement des groupes cohérents qui ne nécessitent pas de redécoupage.*

4.2.4 Étape 3 : Évaluation de la significativité

Une fois le regroupement optimal trouvé pour chaque prédicteur, on calcule le χ^2 global pour la table de contingence réduite : χ^2_{optimal} = statistique pour la table optimale avec $(g-1)(d-1)$ degrés de liberté, où g est le nombre final de groupes.

La p-value associée mesure la force de l'association entre le prédicteur optimisé et la variable dépendante.

4.2.5 Étape 4 : Sélection du meilleur prédicteur

Parmi tous les prédicteurs optimisés, on sélectionne celui qui maximise l'association globale avec la variable dépendante.

1. **Critère :** On retient le prédicteur ayant la p-value la plus faible (ou le χ^2 le plus élevé).
2. C'est cette variable qui sépare le mieux les données à cette étape.
3. Ce prédicteur n'est validé que si sa p-value est inférieure à α_{split} (seuil de partition, typiquement 0.05).

Si aucun prédicteur ne satisfait le critère, le nœud devient une feuille (nœud terminal).

4.2.6 Étape 5 : Partition et récursion

Les données sont partitionnées selon les groupes optimaux du prédicteur sélectionné, créant g nœuds fils.

Pour chaque nœud fils :

— Si le nœud contient moins de $n_{\min}$ observations (typiquement 50-100), il devient une feuille

— Si la profondeur maximale est atteinte, il devient une feuille

— Sinon, on répète les étapes 1 à 5 récursivement

4.3 Heuristique versus optimum global

4.3.1 Complexité du problème optimal

Trouver la partition **vraiment** optimale nécessiterait :

1. D'énumérer toutes les partitions possibles en g groupes pour $g = 2, \ldots, c$
2. De calculer le χ^2 pour chacune
3. De sélectionner la partition avec le χ^2 maximal (p-value minimale)

Complexité computationnelle.

— **Prédicteurs nominaux** : $B(c)$ partitions (croissance super-exponentielle)

— **Prédicteurs ordinaux** : 2^{c-1} partitions (croissance exponentielle)

Exemple 4. *Pour $c = 8$ catégories nominales : $B(8) = 4140$ partitions à examiner.*

Cette énumération devient rapidement impraticable pour plusieurs prédicteurs et plusieurs nœuds.

4.3.2 Justification de l'heuristique

L'heuristique de fusion proposée par Kass :

— A une complexité de $O(c)$ pour les prédicteurs ordinaux

— A une complexité de $O(c^2)$ pour les prédicteurs nominaux

— Fournit des résultats très satisfaisants en pratique

— Ne garantit pas l'optimum global mais s'en approche

4.4 CHAID avec ajustement de Bonferroni

4.4.1 Problème des tests multiples

Inflation du risque de type I. Lorsqu'on effectue m tests d'hypothèses indépendants, chacun au niveau α, la probabilité de commettre au moins une erreur de type I (trouver une différence qui n'existe pas) augmente mécaniquement avec le nombre de tests. C'est le **risque d'inflation du risque alpha**.

La probabilité de commettre au moins un rejet erroné est donnée par :

$$P(\text{au moins un rejet erroné}) = 1 - (1 - \alpha)^m$$

Le tableau suivant met en évidence l'augmentation du risque alpha global en fonction du nombre de tests (pour un seuil initial $\alpha = 0.05$) :

Nombre de tests (m)	**Risque global d'erreur ($1 - 0,95^m$)**
1	5 %
2	$\approx$ 10 %
10	$\approx$ 40 %
50	$\approx$ 92 %

Exemple 5. *Avec $m = 20$ tests indépendants à $\alpha = 0.05$, on a environ 64 % de chances de rejeter au moins une hypothèse nulle vraie. Autrement dit, plus on cherche, plus on a de chances de trouver une corrélation par pur hasard.*

Application à CHAID. Dans l'algorithme CHAID, la multiplicité des tests est inhérente car :

— Nous examinons toutes les paires de catégories lors de la fusion ;

— Nous comparons différentes partitions possibles ;

— Nous testons plusieurs prédicteurs simultanément.

Sans correction, l'arbre risquerait de devenir trop complexe en détectant des associations « significatives » qui ne sont que du bruit statistique.

4.4.2 Principe de la correction de Bonferroni

Formule générale. Pour contrer cette inflation, la correction de Bonferroni ajuste le seuil de signification individuel. Si nous effectuons m tests, le nouveau seuil α' est calculé comme suit :

$$\alpha' = \frac{\alpha}{m}$$

Ou de manière équivalente, on multiplie la p-value observée par m : $p_{\text{ajustée}} = \min(m \times p, 1)$.

Le tableau ci-dessous illustre la sévérité de la correction pour un α initial de 5 % :

Nombre de tests (m)	Seuil Bonferroni corrigé (α/m)
1	5,00 %
2	2,50 %
3	1,67 %
4	1,25 %
5	1,00 %
10	0,50 %
20	0,25 %

Justification théorique. Sous l'hypothèse d'indépendance des tests (ou par l'inégalité de Boole), cette correction garantit que le risque global reste sous contrôle :

$$P(\text{au moins un rejet erroné}) \leq m \times \frac{\alpha}{m} = \alpha$$

4.4.3 Application dans CHAID : multiplicateur de Bonferroni

Kass (1980) propose d'ajuster la p-value du χ^2 optimal par un multiplicateur m spécifique, égal au nombre de façons possibles de partitionner c catégories en g groupes.

Formules selon le type de prédicteur. Prédicteurs nominaux :

$$m_{\text{nominal}}(c, g) = S(c, g) = \sum_{i=0}^{g-1} (-1)^i \frac{(g-i)^c}{i!(g-i)!}$$

(Nombre de Stirling de seconde espèce multiplié par $g!$)

Prédicteurs ordinaux :

$$m_{\text{ordinal}}(c, g) = \binom{c-1}{g-1}$$

Prédicteurs flottants :

$$m_{\text{flottant}}(c, g) = \binom{c-2}{g-2} + g\binom{c-2}{g-1}$$

Exemple 6. *Pour un prédicteur nominal avec $c = 8$ catégories regroupé en $g = 3$ groupes, le multiplicateur est $m = 966$. La p-value observée devra être extrêmement faible pour rester significative après avoir été multipliée par 966.*

4.4.4 Impacts, Avantages et Limites

L'utilisation de la correction de Bonferroni dans CHAID et en statistiques générales présente une balance entre rigueur et sensibilité.

Avantages : Lutte contre le « Data Dredging ».

- **Contrôle de l'erreur :** Elle corrige efficacement le risque d'inflation du risque alpha global.
- **Intégrité scientifique :** Elle limite le risque de *data dredging* (triturage de données). Cette pratique consiste à réaliser un grand nombre de tests sur de nombreuses variables et à ne publier ou ne retenir que celles qui sont significatives (aussi appelé *p-hacking*). Bonferroni agit comme un garde-fou contre les découvertes fortuites.

Inconvénients : Perte de Puissance.

— **Conservatisme :** La correction est souvent considérée comme conservatrice (trop stricte), surtout lorsque les tests sont corrélés entre eux.

— **Diminution de la puissance :** En abaissant drastiquement le seuil α, on augmente le risque de type II (ne pas rejeter H0 alors qu'elle est fausse). On risque ainsi de passer à côté d'effets réels mais faibles.

Conséquence spécifique pour CHAID. Cela se traduit par une pénalisation des prédicteurs ayant un grand nombre de catégories (c). Plus c est grand, plus le multiplicateur m est élevé. Cela compense le biais naturel des arbres de décision qui favorisent habituellement les variables très ramifiées.

Changement de structure d'arbre. La correction de Bonferroni peut :

— Changer l'ordre de sélection des prédicteurs

— Réduire la profondeur de l'arbre (en rendant certains splits non significatifs)

— Favoriser des prédicteurs plus simples

4.5 Règles d'arrêt

CHAID utilise plusieurs critères pour décider quand arrêter la croissance de l'arbre :

1. **Seuil de significativité** $\alpha_{\textbf{split}}$: Ne pas partitionner si $p_{\text{ajustée}} > \alpha_{\text{split}}$
2. **Profondeur maximale** $L_{\textbf{max}}$: Limiter la profondeur de l'arbre (typiquement 5-10 niveaux)
3. **Taille minimale du nœud parent** $n_{\textbf{parent}}$: Ne pas tenter de partitionner un nœud avec moins de n_{parent} observations (typiquement 100)
4. **Taille minimale du nœud fils** $n_{\textbf{fils}}$: Ne considérer que les partitions où chaque groupe a au moins n_{fils} observations (typiquement 50)

Remarque 7. *Ces règles préviennent le sur-ajustement et assurent que les divisions sont basées sur des échantillons suffisamment grands pour être statistiquement fiables.*

5 Exemple d'application : Prédiction de la réussite scolaire

5.1 Description des données

Pour illustrer l'algorithme CHAID, nous utilisons le jeu de données *Student Performance* disponible sur le dépôt UCI Machine Learning Repository. Ce dataset contient des informations sur 649 étudiants portugais et comprend diverses variables socio-démographiques, comportementales et scolaires.

Variable cible. Nous avons créé une variable binaire `outcome` à partir de la note finale (G3) :

— **Pass** : note ≥ 10 (549 étudiants, 84.6%)

— **Fail** : note < 10 (100 étudiants, 15.4%)

Variables explicatives. Nous avons sélectionné 7 prédicteurs représentant les trois types de variables :

Variable	Type	Description
`failures_cat`	ORDINAL	Nombre d'échecs passés : $0 < 1 < 2 < 3+$
`study_time`	ORDINAL	Temps d'étude hebdomadaire : <2h < 2-5h < 5-10h < >10h
`mother_education`	ORDINAL	Niveau d'éducation : None < Primary < Secondary < Higher
`absence_level`	FLOATING	Niveau d'absences : None < Low < Medium < High + *miss*
`sex`	NOMINAL	Sexe de l'étudiant : M, F
`higher`	NOMINAL	Aspiration aux études supérieures : yes, no
`internet`	NOMINAL	Accès internet à domicile : yes, no

TABLE 1 – Variables explicatives et leurs types

Traitement des valeurs manquantes. Pour la variable `absence_level`, nous avons artificiellement introduit 10% de valeurs manquantes (étiquetées *miss*) pour illustrer le fonctionnement d'un prédicteur flottant. Cette catégorie spéciale peut se fusionner librement avec n'importe quel groupe d'absences.

Paramètres de l'algorithme. Nous avons utilisé les paramètres suivants :

- $\alpha_{\text{fusion}} = 0.05$: seuil pour fusionner les catégories
- $\alpha_{\text{split}} = 0.05$: seuil pour créer une partition
- Profondeur maximale : 3 niveaux
- Taille minimale du nœud parent : 30 observations
- Taille minimale du nœud fils : 15 observations

5.2 Processus de fusion selon le type de prédicteur

Nous illustrons maintenant comment l'étape de fusion opère différemment selon le type de prédicteur.

5.2.1 Prédicteur ordinal : `failures_cat`

Pour le prédicteur ordinal `failures_cat` ($0 < 1 < 2 < 3+$), seules les paires adjacentes sont testées. Le tableau initial des χ^2 par paire montre :

	0	**1**	**2**	**3+**
0	—	70.94	—	—
1	0.0000	—	0.10	—
2	—	0.7565	—	0.62
3+	—	—	0.4308	—

TABLE 2 – χ^2 par paire pour `failures_cat` au Nœud 0 (triangle supérieur : χ^2, triangle inférieur : p-value). Seules les paires adjacentes sont calculées.

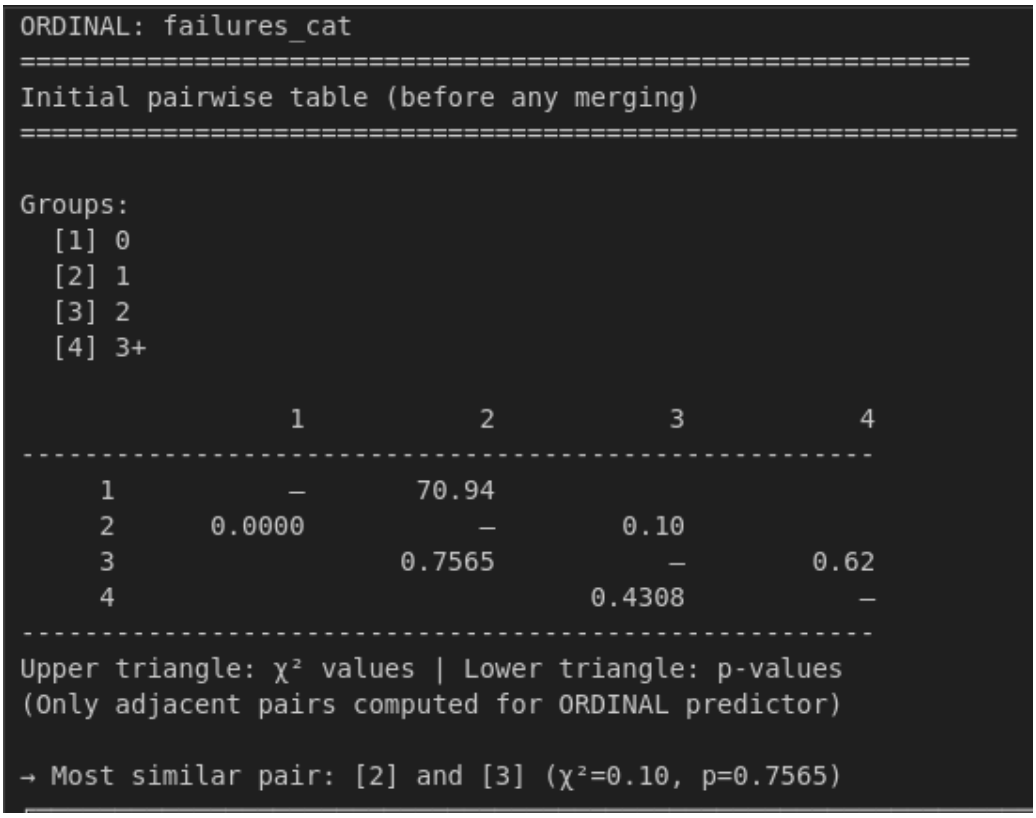


FIGURE 1 – Tableau des paires pour `failures_cat` au Nœud 0 — Sortie Python

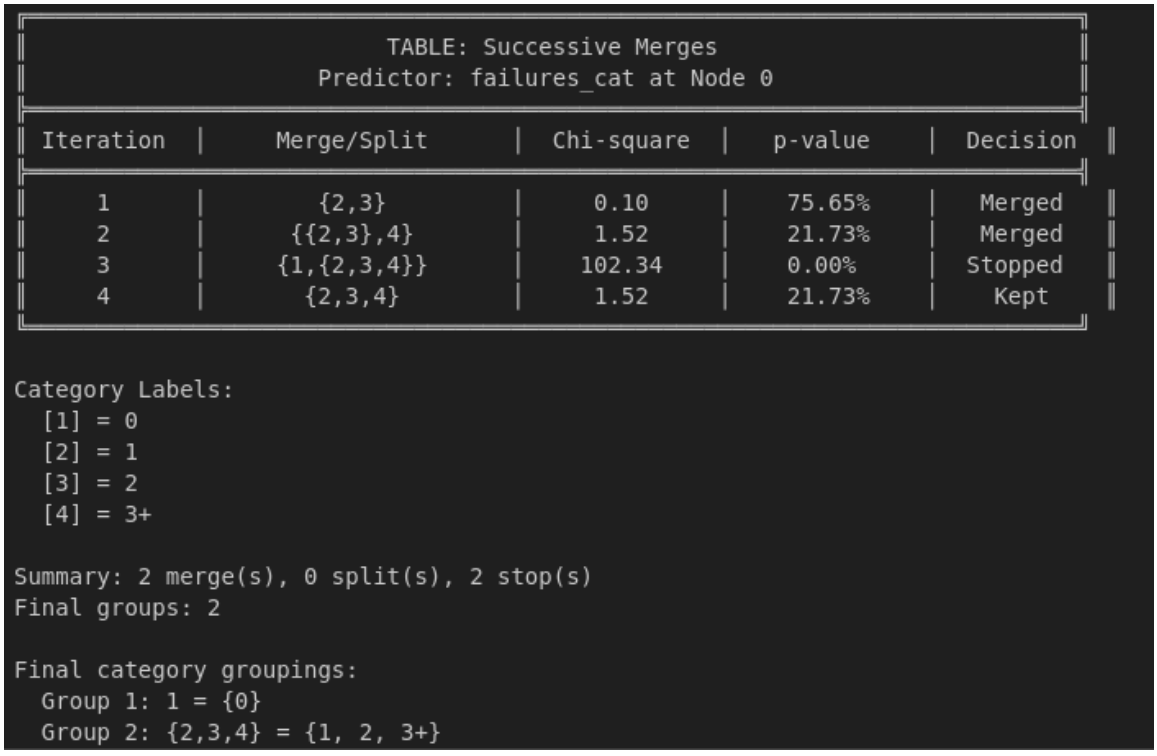


FIGURE 2 – Fusions successives pour `failures_cat` au Nœud 0 — Itérations 1 à 4

Interprétation des fusions successives :

1. **Itération 1** : La paire $\{1, 2\}$ a le χ^2 le plus faible (0.10, p=0.7565 $>$ 0.05). Ces catégories sont **fusionnées** car non significativement différentes.
2. **Itération 2** : On teste ensuite $\{\{1, 2\}, 3+\}$, obtenant $\chi^2 = 1.52$ (p=0.2173 $>$ 0.05). Ces groupes sont également **fusionnés**.
3. **Itération 3** : Reste à tester $\{0, \{1, 2, 3+\}\}$, avec $\chi^2 = 102.34$ (p $\approx$ 0). Cette différence est hautement significative, donc la fusion s'arrête.
4. **Vérification (splitting)** : Le groupe $\{1, 2, 3+\}$ contient 3 catégories. On vérifie s'il doit être re-divisé. Aucune division n'améliore significativement le modèle, donc le regroupement est conservé.

Résultat final : Deux groupes sont retenus :

— Groupe 1 : $\{0\}$ — étudiants sans échec passé (431 obs., 90.7% de réussite)

— Groupe 2 : $\{1, 2, 3+\}$ — étudiants avec au moins un échec (218 obs., 72.9% de réussite)

5.2.2 Prédicteur nominal : `higher`

Pour un prédicteur nominal comme `higher` (yes/no), toutes les paires peuvent être testées :

	no	**yes**
no	—	62.25
yes	0.0000	—

TABLE 3 – χ^2 par paire pour `higher` au Nœud 0 (triangle supérieur : χ^2, triangle inférieur : p-value)

```
NOMINAL: higher
============================================================
Initial pairwise table (before any merging)
====================================

Groups:
  [1] no
  [2] yes

                1          2
------------------------------
     1          –      62.25
     2     0.0000          –
------------------------------
Upper triangle: χ² values | Lower triangle: p-values

→ Most similar pair: [1] and [2] (χ²=62.25, p=0.0000)
```

FIGURE 3 – Tableau des paires pour `higher` au Nœud 0 — Sortie Python

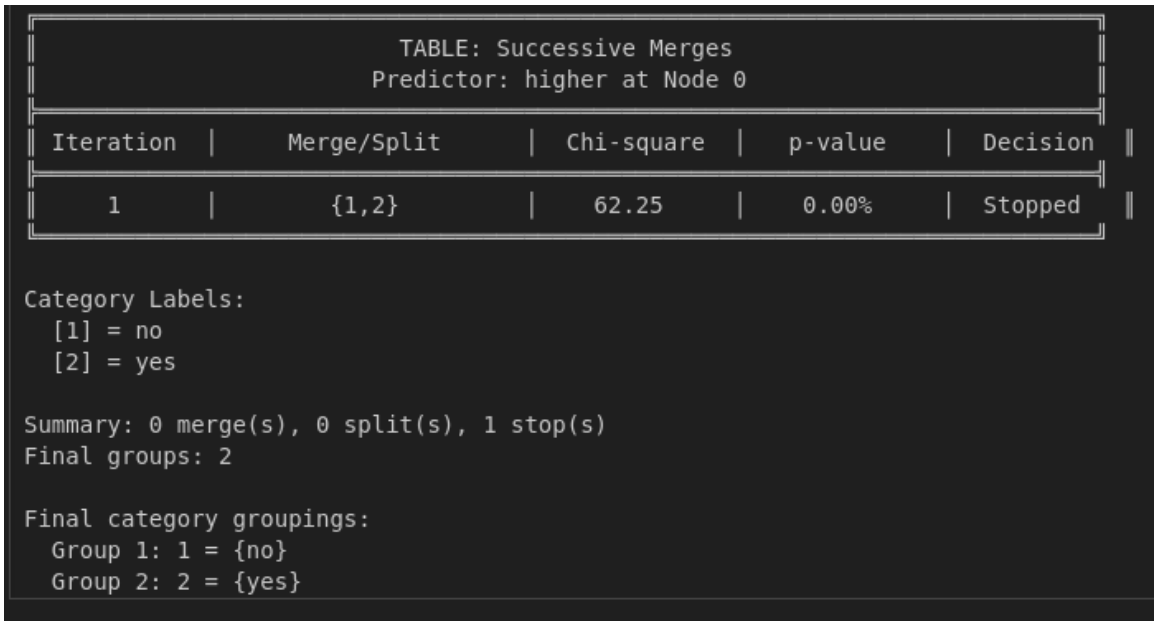


FIGURE 4 – Test de fusion pour `higher` au Nœud 0 — Itération 1

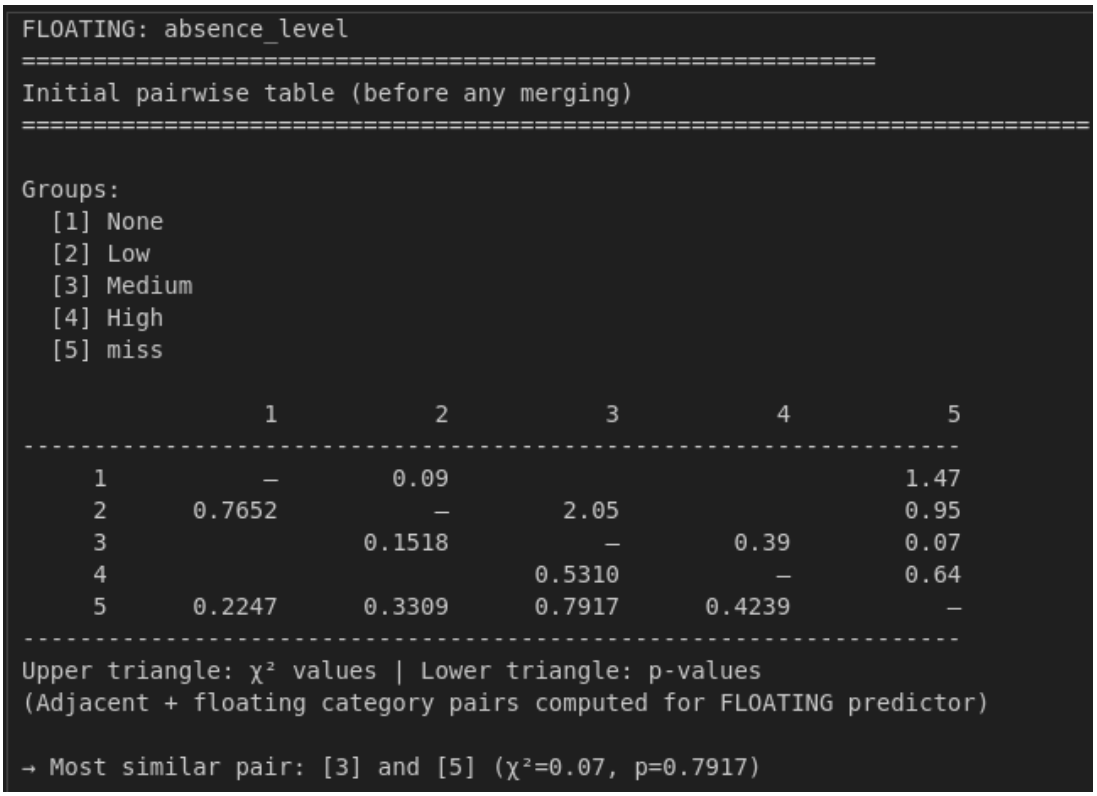
Le χ^2 de 62.25 ($p \approx 0$) indique une différence hautement significative entre les deux catégories. Aucune fusion n'est effectuée. Les étudiants aspirant aux études supérieures réussissent à 87.8%, contre seulement 63.8% pour les autres.

5.2.3 Prédicteur flottant : `absence_level`

Pour le prédicteur flottant `absence_level`, les catégories ordinales (None < Low < Medium < High) suivent la contrainte de contiguïté, mais la catégorie flottante *miss* peut se combiner avec n'importe quelle autre :

	None	**Low**	**Medium**	**High**	**miss**
None	—	0.09	—	—	1.47
Low	0.7652	—	2.05	—	0.95
Medium	—	0.1518	—	0.39	0.07
High	—	—	0.5310	—	0.64
miss	0.2247	0.3309	0.7917	0.4239	—

TABLE 4 – χ^2 par paire pour `absence_level` au Nœud 0 (triangle supérieur : χ^2, triangle inférieur : p-value). La ligne/colonne *miss* montre toutes les comparaisons possibles (catégorie flottante).



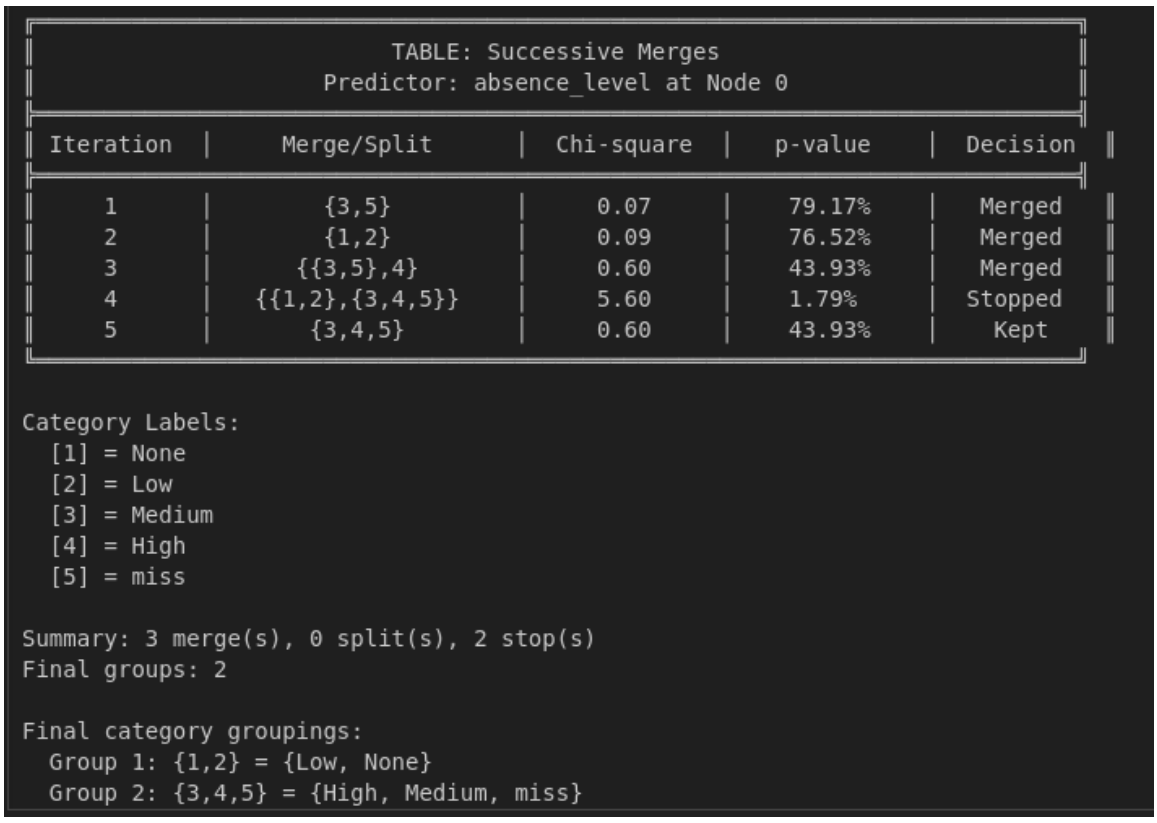
```
FLOATING: absence_level
============================================================
Initial pairwise table (before any merging)
================================================================================

Groups:
  [1] None
  [2] Low
  [3] Medium
  [4] High
  [5] miss

                1          2          3          4          5
--------------------------------------------------------------------
     1          –       0.09                                1.47
     2     0.7652          –       2.05                     0.95
     3                  0.1518          –       0.39        0.07
     4                               0.5310          –       0.64
     5     0.2247     0.3309     0.7917     0.4239          –
--------------------------------------------------------------------
Upper triangle: χ² values | Lower triangle: p-values
(Adjacent + floating category pairs computed for FLOATING predictor)

→ Most similar pair: [3] and [5] (χ²=0.07, p=0.7917)
```

FIGURE 5 – Tableau des paires pour `absence_level` au Nœud 0 — Sortie Python



```
╔═══════════════════════════════════════════════════════════════════════╗
║                        TABLE: Successive Merges                       ║
║                    Predictor: absence_level at Node 0                 ║
╠═══════════╦═══════════════════╦═══════════════╦═══════════════╦═══════╗
║ Iteration ║    Merge/Split    ║  Chi-square   ║    p-value    ║ Decision ║
╠═══════════╬═══════════════════╬═══════════════╬═══════════════╬═══════╣
║     1     ║       {3,5}       ║     0.07      ║    79.17%     ║  Merged  ║
║     2     ║       {1,2}       ║     0.09      ║    76.52%     ║  Merged  ║
║     3     ║     {{3,5},4}     ║     0.60      ║    43.93%     ║  Merged  ║
║     4     ║ {{1,2},{3,4,5}}   ║     5.60      ║    1.79%      ║ Stopped  ║
║     5     ║     {3,4,5}       ║     0.60      ║    43.93%     ║   Kept   ║
╚═══════════╩═══════════════════╩═══════════════╩═══════════════╩═══════╝

Category Labels:
  [1] = None
  [2] = Low
  [3] = Medium
  [4] = High
  [5] = miss

Summary: 3 merge(s), 0 split(s), 2 stop(s)
Final groups: 2

Final category groupings:
  Group 1: {1,2} = {Low, None}
  Group 2: {3,4,5} = {High, Medium, miss}
```

FIGURE 6 – Fusions successives pour `absence_level` au Nœud 0 — Itérations 1 à 5

Fusions successives :

1. **Itération 1** : $\{\text{Medium}, \text{miss}\}$ ont le χ^2 le plus faible (0.07, p=0.7917). Fusion effectuée.
2. **Itération 2** : $\{\text{None}, \text{Low}\}$ (0.09, p=0.7652). Fusion effectuée.
3. **Itération 3** : $\{\{\text{Medium}, \text{miss}\}, \text{High}\}$ (0.60, p=0.4393). Fusion effectuée.
4. **Arrêt** : $\{\{\text{None}, \text{Low}\}, \{\text{Medium}, \text{High}, \text{miss}\}\}$ montre un $\chi^2 = 5.60$ (p=0.0179), mais après correction de Bonferroni, p=0.1255 > 0.05.

Résultat final : Deux groupes :

— Groupe 1 : $\{\text{None}, \text{Low}\}$ — absences faibles/nulles (412 obs., 87.6% réussite)

— Groupe 2 : $\{\text{Medium}, \text{High}, \text{miss}\}$ — absences élevées ou manquantes (237 obs., 76.8% réussite)

Remarquons que la catégorie *miss* s'est naturellement associée aux niveaux d'absences élevés, suggérant que les données manquantes pourraient être liées à un comportement d'absentéisme.

5.3 Sélection du meilleur prédicteur

Une fois tous les prédicteurs optimisés par fusion/division, l'algorithme calcule le χ^2 global pour chacun et sélectionne celui ayant la p-value la plus faible :

Prédicteur	**Type**	**#cat**	**#grp**	χ^2	**df**	**p-value**
`failures_cat`	ORDINAL	4	2	102.34	1	0.0000 ⋆
`higher`	NOMINAL	2	2	62.25	1	0.0000
`study_time`	ORDINAL	4	3	19.25	2	0.0002
`mother_education`	ORDINAL	4	2	10.09	1	0.0045
`internet`	NOMINAL	2	2	5.05	1	0.0246
`sex`	NOMINAL	2	2	3.97	1	0.0463
`absence_level`	FLOATING	5	2	5.60	1	0.1255

TABLE 5 – Comparaison des prédicteurs optimisés au Nœud 0 (premier niveau de l'arbre). Le symbole ⋆ indique le prédicteur sélectionné.

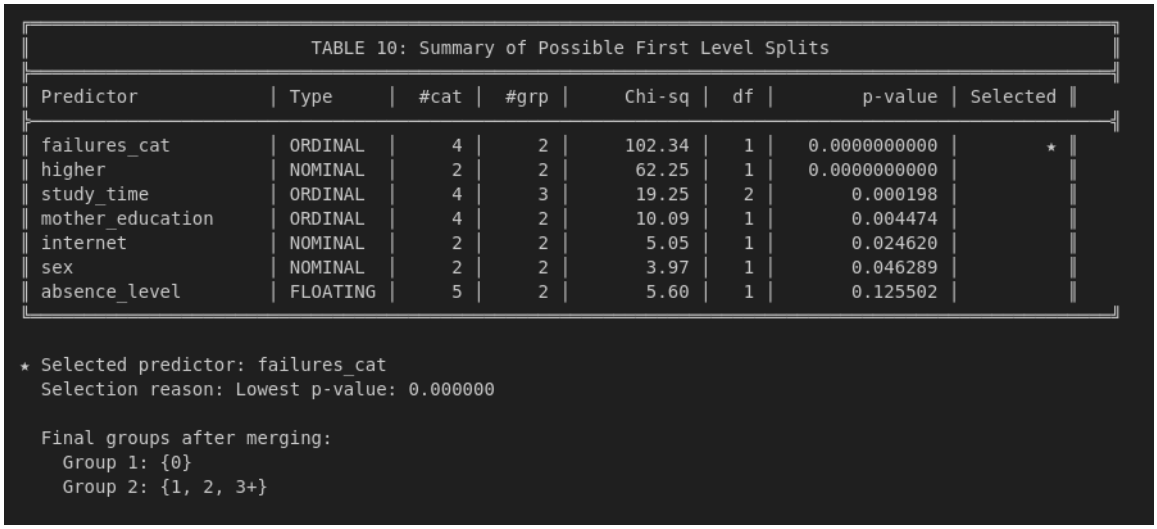


FIGURE 7 – Comparaison des prédicteurs au Nœud 0 — Sortie Python

Interprétation. Le prédicteur `failures_cat` est sélectionné avec un $\chi^2 = 102.34$, largement supérieur aux autres. Les échecs scolaires passés sont donc le facteur le plus discriminant pour prédire la réussite. Après correction de Bonferroni, la p-value reste hautement significative.

5.4 Arbre de décision final

L'algorithme CHAID a construit un arbre de profondeur 3 avec 7 nœuds (dont 4 feuilles). La structure finale est présentée dans la Figure 8.

Lecture de l'arbre.

- **Racine (Nœud 0)** : Tous les étudiants (n=649, 84.6% Pass)
  - Division selon `failures_cat` ($\chi^2 = 102.34$, p≈0)
- **Branche gauche — Aucun échec passé (Nœud 1)** : n=431, 90.7% Pass
  - Division selon `higher` ($\chi^2 = 40.96$, p≈0)
  - **Nœud 2** (higher=no) : n=36, 61.1% Pass
  - **Nœud 3** (higher=yes) : n=395, 92.7% Pass — Meilleur profil
- **Branche droite — Au moins un échec (Nœud 6)** : n=218, 51.0% Pass
  - Division selon `study_time` ($\chi^2 = 6.97$, p=0.0248)
  - **Nœud 4** (study_time ∈ {<2h, <2h}) : n=90, 38.9% Pass — Profil à risque élevé
  - **Nœud 5** (study_time ∈ {5-10h, >10h}) : n=128, 98.3% Pass

Insights principaux.

1. **Les échecs passés sont le facteur le plus important** : La première division sépare clairement les étudiants sans échec (90.7% réussite) de ceux ayant échoué au moins une fois (51.0%).
2. **L'aspiration joue un rôle protecteur** : Parmi les étudiants sans échec, ceux aspirant aux études supérieures réussissent à 92.7%, contre 61.1% pour les autres.
3. **Le temps d'étude compense fortement les échecs** : Pour les étudiants avec échecs passés, ceux qui étudient au moins 5h/semaine ont un taux de réussite exceptionnel de 98.3%, contre seulement 38.9% pour ceux qui étudient moins de 2h.
4. **Les absences n'apparaissent pas dans l'arbre** : Bien que `absence_level` montre une association avec la réussite (p=0.0179 avant correction), elle est supplantée par des prédicteurs plus puissants après correction de Bonferroni.

5.5 Comparaison avec et sans correction de Bonferroni

Pour évaluer l'impact de la correction de Bonferroni, nous avons construit deux arbres : un avec ajustement (Figure 9) et un sans ajustement (Figure 10).

Résultat : arbres identiques. Les deux arbres produisent exactement la même structure. Ce résultat, loin d'être problématique, révèle une caractéristique importante de nos données : **les associations sont suffisamment fortes pour résister à la correction de Bonferroni**.

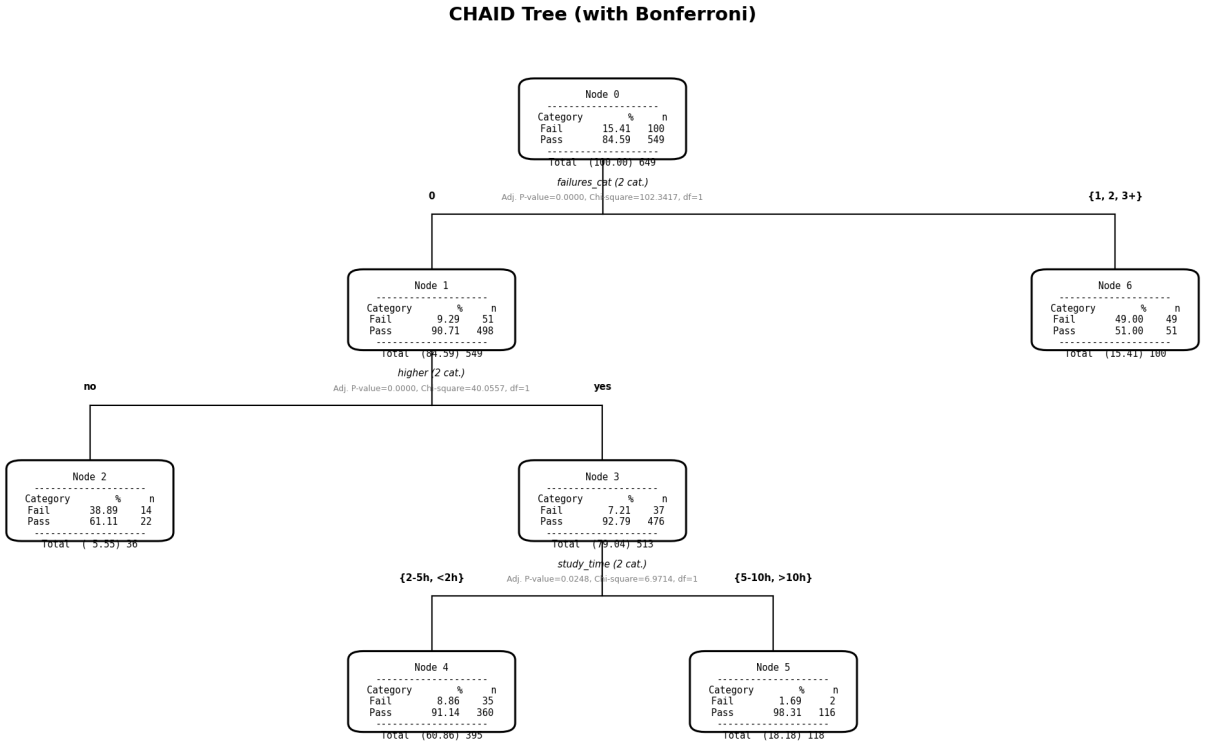


FIGURE 8 – Arbre CHAID final pour la prédiction de réussite scolaire (avec correction de Bonferroni)

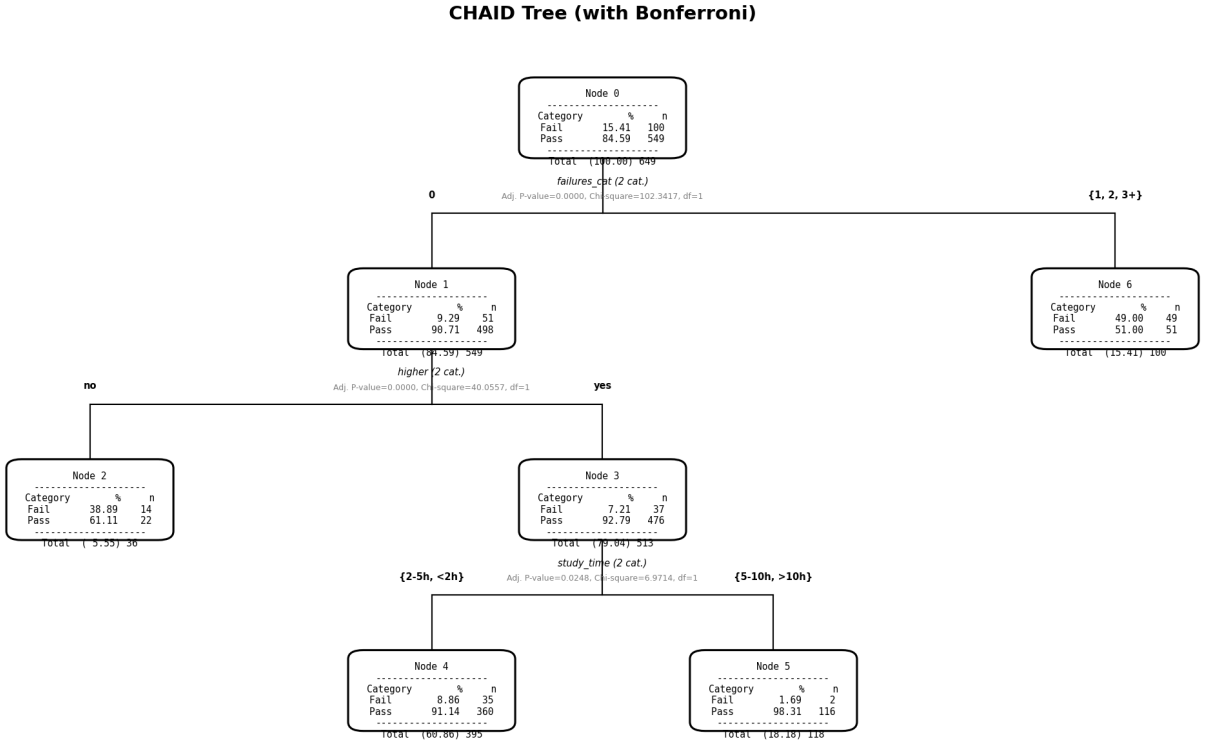


FIGURE 9 – Arbre CHAID avec correction de Bonferroni

Explication. Le tableau suivant compare les p-values avant et après ajustement pour le premier niveau de l'arbre :

Prédicteur	χ^2	p-value	Mult.	p-adj.	Sig. ?
`failures_cat`	102.34	$< 10^{-10}$	1	$< 10^{-10}$	Oui / Oui
`higher`	62.25	$< 10^{-14}$	1	$< 10^{-14}$	Oui / Oui
`study_time`	19.25	0.0001	3	0.0003	Oui / Oui
`mother_education`	10.09	0.0045	2	0.0089	Oui / Oui
`internet`	5.05	0.0246	1	0.0246	Oui / Oui
`sex`	3.97	0.0463	1	0.0463	Oui / Oui
`absence_level`	5.60	0.0179	7	**0.1255**	Oui / Non

TABLE 6 – Impact de la correction de Bonferroni sur les p-values au Nœud 0. La colonne « Mult. » indique le multiplicateur de Bonferroni. La dernière colonne indique la significativité sans/avec correction au seuil $\alpha = 0.05$.

Observations clés.

1. **Prédicteurs sélectionnés inchangés** : Les trois prédicteurs utilisés dans l'arbre (`failures_cat`, `higher`, `study_time`) ont des p-values tellement faibles qu'ils restent hautement significatifs même après correction. Par exemple, `failures_cat` a une p-value pratiquement nulle, et même multipliée par un facteur de Bonferroni, elle reste largement inférieure à 0.05.
2. **Seul `absence_level` est affecté** : Ce prédicteur passe de significatif (p=0.0179) à non significatif (p=0.1255) après correction. Le multiplicateur de Bonferroni pour un prédicteur flottant à 5 catégories formant 2 groupes est :
$$m = \binom{c-2}{g-2} + g\binom{c-2}{g-1} = \binom{3}{0} + 2\binom{3}{1} = 1 + 6 = 7$$
D'où l'ajustement : $0.0179 \times 7 = 0.1255 > 0.05$.
3. **Signal fort vs. bruit statistique** : Le fait que l'arbre soit identique avec ou sans correction suggère que notre modèle capture des effets réels et robustes, non des artefacts statistiques dus à des tests multiples.

Interprétation pratique. Cette invariance à la correction de Bonferroni est un **indicateur de qualité** : nos prédicteurs principaux (échecs passés, aspiration, temps d'étude) ont une relation prédictive forte avec la réussite scolaire. Si nos résultats dépendaient fortement de la correction, cela aurait suggéré des associations faibles ou potentiellement fortuites.

Dans des contextes avec des signaux plus faibles (études exploratoires, données bruitées, nombreux prédicteurs peu informatifs), la correction de Bonferroni aurait un impact beaucoup plus visible en éliminant les divisions non robustes et en réduisant la profondeur de l'arbre.

Cas d'usage typiques de Bonferroni. La correction de Bonferroni devient cruciale dans les situations suivantes :

— **Nombreux prédicteurs faiblement corrélés** : Dans les études de génomique ou d'imagerie médicale, où l'on teste des milliers de variables simultanément, la correction évite de détecter des faux positifs.

— **Exploration de données** : Lorsque l'analyste teste de nombreuses hypothèses sans théorie préalable, la correction protège contre le *p-hacking*.

— **Prédicteurs avec nombreuses catégories** : Un prédicteur nominal avec 10+ catégories aura un multiplicateur de Bonferroni très élevé (plusieurs centaines), rendant difficile la sélection de ce prédicteur sauf si l'association est extrêmement forte.

Dans notre cas, les échecs scolaires ont un pouvoir prédictif si élevé ($\chi^2 > 100$) que la correction ne change rien. Cela confirme la robustesse de nos résultats et la pertinence de ce prédicteur pour comprendre la réussite académique.

Références

— Kass, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29(2), 119-127.

— Ritschard, G. (2013). CHAID and earlier supervised tree methods. In J.J. McArdle & G. Ritschard (eds), *Contemporary Issues in Exploratory Data Mining in Behavioral Sciences* (pp. 48-74). New York : Routledge.

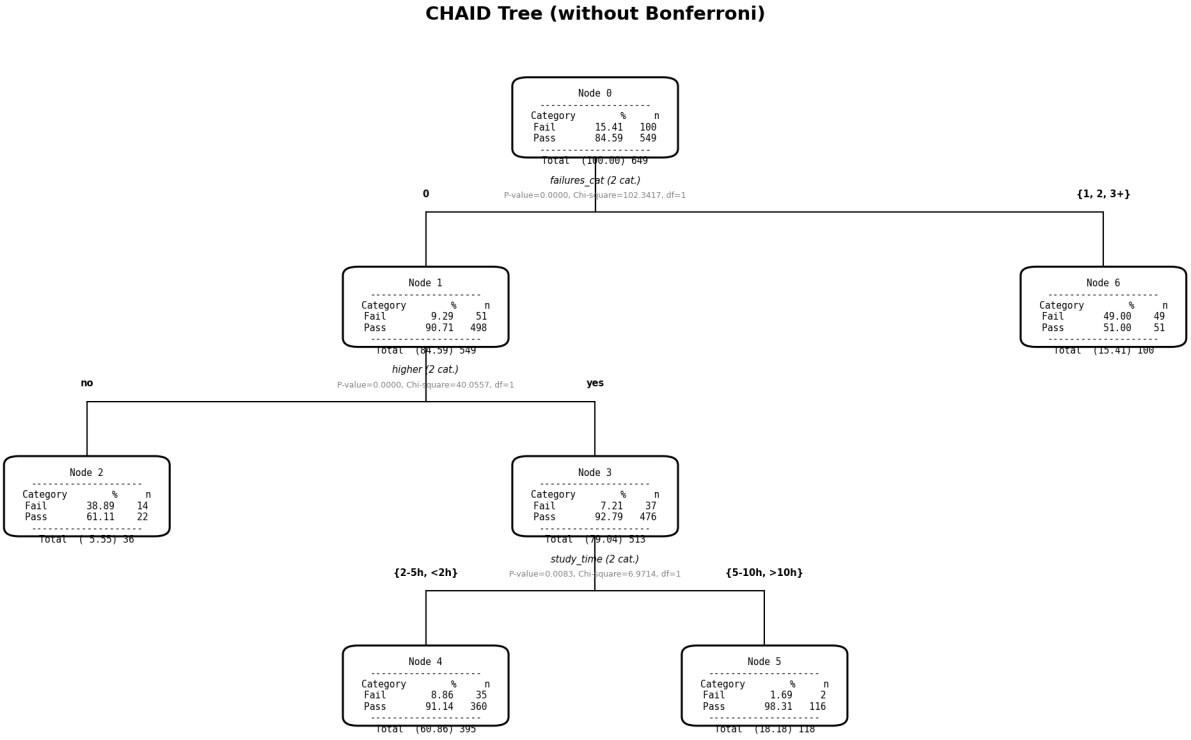


FIGURE 10 – Arbre CHAID sans correction de Bonferroni